自然言語処理論 I

7. 形態素解析(日本語の単語分割)

形態素(morpheme)とは?

- 意味を持つ最小の言語単位
  - 単語よりも小さい単位
- 英語の場合
  - 単語=語幹+接辞
  - play-ing, un-kind-ly
- 日本語の場合
  - 活用語 (食べ+る)
  - 派生語(寒+さ)
  - 複合語(財務+省)

形態素解析

- 構文解析の前に行われる処理
- 主な仕事
  - 形態素区切りを決める ← 英語
    - ✦ playing = play + ing
  - 品詞を決める
  - 単語境界を決める ← 日本語

品詞(part-of-speech)

- 日本語の品詞体系で主に使われる品詞
- 自立語
  - 動詞、形容詞、形容動詞 (活用語)
  - 名詞、副詞、連体詞、接続詞、感動詞
- 付属語
  - 助動詞 (活用語)
  - 助詞、語尾

日本語の形態素解析

- 単語に区切って、品詞を決める
- 例「くるまでまつ」
  - くるま(名詞) で(助詞) ま(動詞) つ(語尾)
  - くる(動詞) まで(助詞) ま(動詞) つ(語尾)
- どのような知識が必要か
  - 単語辞書
  - 連接可能性辞書(接続表)

単語辞書

- 単語のデータベース
- 記載されているべき情報
  - 品詞
  - 読み

単語辞書の例

見出し語	読み	品詞	
こ	コ	接尾語	(個)
こと	コト	名詞:形式名詞	(事)
この	コノ	連体詞	
た	タ	助動詞	
で	デ	助詞:格助詞	
で	デ	動詞語幹:一段	(出る)
と	ト	助詞:格助詞	
と	ト	助詞:接続助詞	

単語辞書の例

見出し語	読み	品詞	
な	ナ	動詞語幹:ラ行五段	(なる)
に	ニ	助詞:格助詞	
にな	ニナ	動詞語幹:ワ行五段	(担う)
ひ	ヒ	名詞:普通名詞	(日)
ひと	ヒト	名詞:普通名詞	(人)
ひとこと	ヒトコト	名詞:普通名詞	(一言)
っ	ッ	語尾	
元気	ゲンキ	名詞:普通名詞	

接続表

- 品詞(または単語)の接続可能性を表した行列
  - 行: 左側の品詞(単語)
  - 列: 右側の品詞(単語)
  - 1: 接続可能，0: 接続不可能
- 制約は緩めに書くべき
  - 連接する可能性のある品詞対(単語対)が1つだけでもあるなら、接続可能にする

接続表の例

	文末	…	名詞:普通名詞	助詞:格助詞	語尾（ら）	語尾（り）	語尾（わ）	語尾（っ）
文頭	0	…	1	0	0	0	0	0
⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮
名詞:普通名詞	1	…	1	1	0	0	0	0
動詞語幹:ラ行五段	0	…	0	0	1	1	0	1
動詞語幹:ワ行五段	0	…	0	0	0	0	1	1

単語ラティス

- 形態素解析結果を表すグラフ構造
  - ノード: 単語と品詞
  - リンク: 連接可能である単語を結ぶ
- 単語辞書、接続表をもとに作成
  - 作成例 → 添付資料

単語ラティス作成アルゴリズム

- 文頭,文末 というノードを用意
- for i=0 to k
  - 位置iで始まる単語を単語辞書で検索し、該当する単語をノードとして追加
  - 位置iで終わるノード(単語)とiで始まるノード(単語)との連接可能性を接続表で調べる
    接続可能なノード間にリンクを張る
  - 1つもリンクを張れなかったノードを削除

解の優先順位付け

- 単語ラティスには複数の解がある
  - 文頭から文末へのパスは全て解
  - どれが正しい解か?
- 解の優先順位付け
  - 形態素解析の解に順位をつける
  - 場合によっては解をひとつだけ選択する

解の優先順位付け

- 辞書や接続表だけでは正解は決められない
  - ex. 井上洋助教授 → 「井上洋助」＋「教授」
「井上洋」＋「助教授」
  - 構文解析、意味解析、文脈解析が必要
- (あえて)解の優先順位付けを行う理由
  - 構文解析の前処理とする場合
    - ✦ 構文解析の入力の数を絞り込む
  - 形態素解析を単独で行う場合
    - ✦ 構文解析・意味解析を必要としない場合
    - ✦ ex. 情報検索におけるキーワード抽出
自立語を取り出すだけでも十分

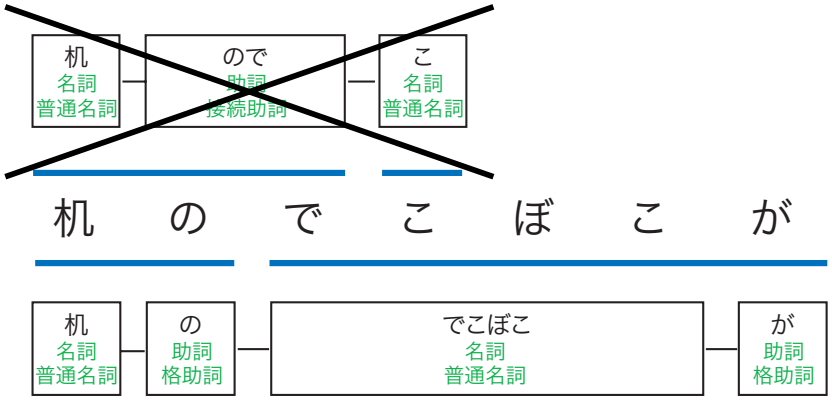
優先規則

- 大きく分けて2種類ある
- 縦型探索型
  - 全ての候補を探索しない
  - 完全な単語ラティスを作らない
- 全解探索型
  - 完全な単語ラティスを作る
  - その中から解を優先的に選択する

優先規則(縦型探索型)

- 最長一致法
  - 長い形態素を優先
- 2文節最長一致法
  - 文節の定義
    - ✦ 1つ以上の自立語と0個以上の付属語を含む単語のグループ
    - ✦ (接頭辞)* (自立語)+ (接尾辞+付属語)*
  - 2文節の長さの和が最長である解を優先

2文節最長一致法



※ 一文節の長さを基準にすると下の解は残らないことに注意

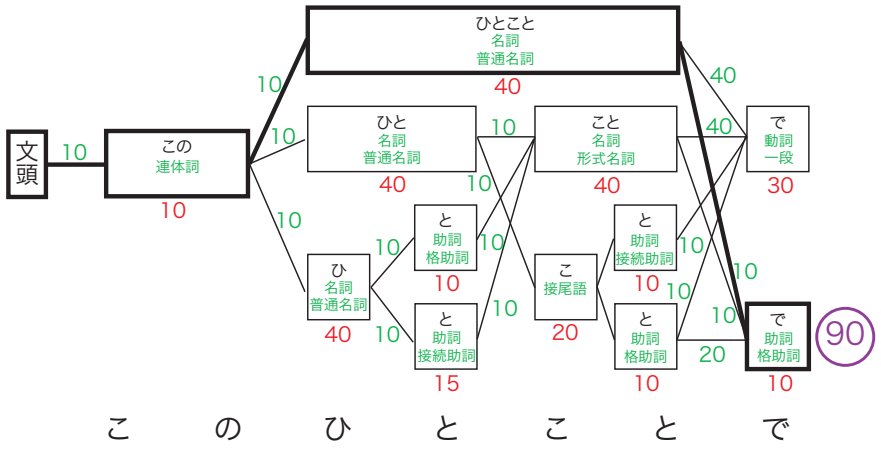
優先規則(全解探索型)

- 形態素数最小法
  - 形態素の数が一番少ない解を優先
- 自立語数最小法
  - 自立語の数が一番少ない解を優先
- 文節数最小法
  - 文節の数が一番少ない解を優先

優先規則(全解探索型)

- コスト最小法
  - 単語とリンクにコストを与える
    - ✦ 良く現れる単語(品詞)ほどコストが低い
    - ✦ 良く連接する単語対(品詞対)ほどコストが低い
  - コストの和が最小になるパスを見つける

コスト最小法



どうやってコストを決めるか?

- 人間が決める
  - 試行錯誤の繰り返し
- 自動的に決める
  - 大量のテキストからコストを学習する
  - よく出てくる単語
    →その単語のコストを低くする
  - よく出てくる品詞対
    →そのリンクのコストを低くする

まとめ

- 日本語の形態素解析
  - 単語の区切り、品詞を決める
- 必要な知識
  - 単語辞書
  - 接続表
- 単語ラティスの作成
- 様々な解の優先順位付け

形態素解析ツール

- フリーのソフトウェア
- JUMAN
  - 京大、東大で開発
  - http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html
- 茶筌
  - 京大、奈良先端大で開発
  - http://chasen.naist.jp/hiki/ChaSen/